Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: DocuSigned by: Zarina R. Bilgrami B31FE2B6CA18416...

Zarina R. Bilgrami
Name

6/22/2023 | 4:45 PM EDT
Date

Title Using Large Language Models to Understand Thought Disorder and Predict Psychosis

Author Zarina R. Bilgrami

Degree Master of Arts

Program Psychology

Approved by the Committee

DocuSigned by:
Phillip M. Wolff
A906AD9BE70D49C...

Phillip M. Wolff
Advisor

DocuSigned by:
Daniel D. Dilks
857692DFF97D455...

Daniel D. Dilks
Committee Member

DocuSigned by:
Elaine Walker
0995D987F17B462...

Elaine Walker
Committee Member

Committee Member

Committee Member

Committee Member

Accepted by the Laney Graduate School:

_____

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

_____

Date

Using Large Language Models to Understand Thought Disorder and Predict Psychosis

By

Zarina R. Bilgrami, B.A., Barnard College, 2018

Advisor: Phillip Wolff, Ph.D.

An abstract of A thesis submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Master of Arts in Psychology
2023

Using Large Language Models to Understand Thought Disorder and Predict Psychosis

By Zarina R. Bilgrami

Thought disorder (TD) is defined as any internal, cognitive disturbance affecting the organization, control, processing, or expression of thought. It is a key characteristic of schizophrenia and psychosis spectrum disorders. Two kinds of TD are widely recognized in schizophrenia and are typically measured through patients' language. One is reflected in local language disorganization, caused by unusual word choices and sequencing. The other concerns non-local disruptions in the flow of ideas across multiple sentences, implying conceptual disorganization. Prior work on TD using computational methods has focused on local disruptions. Here we propose a novel approach to the detection and measurement non-local thought disorganization. We used a new class of artificial intelligence called large language models (LLMs) to examine disorganization across sentences by measuring its ability to predict sentences subjects spoke during their interviews. The model was provided with patient narratives of individuals at clinical high-risk (CHR) for psychosis. The narratives were provided to the model either intact or with the sentence order shuffled. If people's narratives lack global organization, then shuffled narratives should be as effective as intact narratives in at facilitating the prediction of sentences. As expected, intact narratives were more effective than shuffled narratives in facilitating sentence prediction in CHR individuals who did not convert to psychosis (CHR-) than in individuals that converted to psychosis (CHR+). Also as predicted, the LLM was less able to predict individual words in CHR+ than in individuals CHR-. These findings support and expand upon work suggesting that conversion to psychosis is signaled by both local and non-local disruptions in thought.

Using Large Language Models to Understand Thought Disorder and Predict Psychosis

By

Zarina R. Bilgrami
B.A., Barnard College, 2018

A thesis submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Master of Arts in Psychology
2023

Table of Contents

Introduction

Thought disorder (TD) is broadly defined as any internal, cognitive disturbance that affects the organization, control, processing, or expression of thought (Hart and Lewine, 2017). The definition implies that the quality of a person's thoughts may be assessed by measuring how they talk (Andreasen, 1979). TD is considered a key characteristic of schizophrenia and psychosis spectrum disorders and is even apparent in those at risk for developing psychosis (Kuperberg, 2010). In this research, we examine how language data may be used to identify the presence of TD in individuals at high risk for psychosis.

Since its inception, schizophrenia has been linked with two kinds of disruptions, or disorganization, in language. One of these types is signaled by unusual word choices. Elyn Saks provides an example of such local effects from her time in the hospital during a psychotic episode. She was asked if anything was wrong, and she replied:

> *There's cheese and there's whizzes, I'm a cheese whiz. It has to do with effort and subliminal choice. Vertigo and killing.*
>
> Saks, 2007

These sentences are grammatically correct, but the words in them do not often appear together. Moreover, little meaning can be discerned from each of these sentences. This kind of disorganization is local, meaning it occurs between words that are relatively close in proximity and results in word combinations that are relatively hard to understand.

A second kind of TD is signaled by disruptions in language across multiple sentences. For example, one may begin talking to a friend about needing cheese whiz from the store and move on to talking about their love of shopping for handbags, how they helped their grandmother around her nursing home one afternoon, eventually completely losing track of their initial topic of choice: cheez whiz. While all of the sentences in this narrative may be more intelligible than Saks' quote, it is still hard to follow. Disruption in language at the level of sentences is of special interest because it offers

a relatively pure indicator of disorganization at the conceptual level. It offers evidence for disorganization at the conceptual level because languages have few rules specifying acceptable sequences of sentences. Hence, when a sequence of sentences sounds odd, the reason cannot be because of the language itself but rather because of issues in the flow of ideas. Effective sequencing of sentences should preserve cause-and-effect relationships, conceptual categories, and logical relationships. Many have noted how individuals with psychosis may not maintain such conceptual organization (Minor et al., 2014; Minor et al., 2015; Rocca et al., 2018). To date, identifying and measuring non-local disruptions has been possible only through subjective judgments of patient language.

Artificial Intelligence (AI) may offer a more sensitive and reliable approach to identifying disruptions at the conceptual level. Recently, developments in AI have led to a new class of models called large language models (LLMs). These models appear sensitive to the flow of ideas. This is evident in their ability to generate entire essays on a single topic, holding a continuous thread, not just keeping track of the correct ordering of words within sentences. Given these advances, it may be possible to detect whether someone's language includes sequences of sentences that preserve conceptual relationships. Herein we find that such models can indeed track sequences of ideas and that individuals at clinical high risk (CHR) for psychosis differ on the degree to which the sequencing of ideas matters.

Background

TD is present in a range of psychotic disorders and even in their prodromal and risk states. It was central to Kraepelin's (1919) conceptualization of dementia praecox, a precursor to schizophrenia. He noted that psychosis earlier in life (i.e., not age-related dementia) was evidenced in speech by "derailments" and "incoherence" of thought processes. Bleuler (1911), who coined the term "schizophrenia" also gleaned disturbances in mental functions of his patients based on

language. Like Kraepelin, he noticed two related but somewhat different manifestations of what he called, "loosening of associations". One of the processes is exemplified when "associations do not become entirely senseless, but they still appear odd, bizarre, or distorted" (p. 19) and the other when "the most important determinant of associations is lacking the concept of purpose" (Bleuler, 1911 p. 15). While based almost exclusively on observation, these behaviors and terms represent both a local incoherence in which what people say is nonsensical and a broader incoherence (derailment) that makes them hard to follow over the course of a narrative.

Presently, TD is still largely assessed through manual clinician ratings. Diagnostic measures have assessments of TD and communicative disturbance in both positive and negative symptom domains (i.e., Structured Interview for Psychosis-Risk Syndromes (SIPS; Woods, 2019), and Positive and Negative Syndrome Scale (PANSS; Kay, Fiszbein & Opler, 1987). Most often, these are measured using ordinal scales to capture the presence and basic intensity of the disturbance but are not concerned with the heterogeneity in the *nature* of TD that can be seen amongst tested individuals.

There are a number of specific scales and measures for assessing TD, including the Scale for the Assessment of Positive Symptoms (SAPS; Andreasen, 1983), Scale for the Assessment of Negative Symptoms (SANS; Andreasen, 1983), Thought and Language Index (TLI; Liddle et al., 2002), Thought Disorder Index (TDI; Johnson & Holzman, 1979), and Communication Disturbances Scale (CDI; Gordinier & Docherty, 2001). The deepest and most nuanced measure for assessing TD is the Scale for the Assessment of Thought, Language, and Communication (TLC) developed by Andreasen (1979a; 1979b; 1986). The TLC uses three conceptual categories of TD (thought, language, and communication), and each further breaks down into subtypes of TD (of which there are 18 total across the categories). Andreasen (1986) notes that a few of these subtypes together comprise the original concept of "loosening of associations" proposed by Bleuler (1911).

One of these subtypes is "incoherence", and it is defined as a pattern of speech that is largely incomprehensible due to abnormality of word choice and occurs within the level of a sentence or clause. The second relevant subtype is derailment, which is illustrated by a speaker's ideas slipping from one topic to another that is either clearly or obliquely related or completely unrelated. In other words, people's ideas "may shift idiosyncratically from one frame of reference to another" (Andreasen, 1986). Sometimes there may be a vague connection between ideas, while other times people's ideas seem disjointed. It should also be noted that tangentiality, another subtype of TD in the TLC, is a form of derailment that is reserved for loosening of associations that occurs in response to a question. Someone's response to a question may seem related in a distant way, unrelated or totally irrelevant to their interlocuter's question. Several decades after Kraepelin (1919) and Bleuler (1911), the TLC preserves the same categories of both local (incoherence) and non-local (derailment, tangentiality) loosening of associations and further segments them into unique categories of measurement.

Measurement of local disturbance has been the basis for a number of machine learning classifiers. These use natural language processing (NLP) techniques to predict the development of psychosis in CHR. Using methods like latent semantic analysis (LSA), these approaches measure TD in speech as low levels of semantic similarity between adjacent words and sentences (Elvevåg, Foltz, Weinberger, & Goldberg, 2007; Elvevåg, Foltz, Rosenstein, & DeLisi, 2010; Bedi et al., 2015; Corcoran et al., 2018). However, this may not be the complete picture of semantic coherence. High levels of semantic similarity need not imply coherence, in fact they might reflect redundancy, which is not "healthy" speech that one would expect in conversation. Further, low levels of similarity need not indicate disorganization. Adjacent sentences often lack semantic overlap, but they are nevertheless connected by means of entailment relations. For example: *The spider fell.* followed by *The girl jumped up.* sounds natural to our ear, but the two sentences are not especially similar in meaning.

In the following set of experiments, we propose a new approach to the assessment of both local and non-local coherence. We do so by taking advantage of the generative properties of LLMs which allow them to predict different aspects of language. In order to measure local disorganization of ideas, we took advantage of LLMs' ability to predict missing words. To measure non-local disorganization, we took advantage of LLMs' ability to generate sentences. We achieved this by capitalizing on LLMs' ability to track ideas across large swaths of text, which is made possible by their sensitivity to context. We used these features to investigate aspects of TD in speech from CHR at their baseline clinical interview, knowing their eventual diagnostic outcome (i.e., whether they developed a threshold psychotic disorder (CHR+) or not (CHR-) after their baseline interview).

While there are clear limitations, classifiers measuring local measurements of incoherence are effective at predicting psychosis onset in CHR. Therefore, we propose that word-level disorganization might be a particularly useful signature of conversion to psychosis in at-risk youth. For this reason, we predict that LLMs should be better at predicting missing words in transcripts of CHR- individuals than CHR+ individuals. However, if we do find this to be true, further work needs to be done to understand why. It could be due to the fact that they use more similar words from sentence to sentence or even that their sequencing of words is more predictable than those who develop psychosis. However, word-level disruptions do not represent the entirety of TD in psychosis spectrum disorders. In fact, they represent a small portion of the disturbances observed in this population (Andreasen, 1986). Disruptions at the non-local level, which involve the organization of ideas, are a pervasive affliction in CHR and might be more relevant to CHR- than word-level disruptions.

Disruptions in ideas can be detected using LLM's ability to predict entire sentences. This is possible through LLMs' sensitivity to the broader linguistic context, which strongly influences their generative capabilities (Shi & Wolff, 2021). For example, in response to the sentence *A man dashed*

into the store, an LLM might generate *He was in a hurry to buy a last-minute gift for his wife*, or it might produce *He needed a place to hide*. The first sentence is more likely if the context preceding the sentence is about birthdays, and the second is more likely when preceded by a context about bank robberies. The ability to generate sentences and sensitivity to context can be used to detect disruptions in the flow of ideas. Specifically, an LLM can be queried about how a person would respond to questions posed by an interviewer. It is expected that the LLM's performance in this task will improve if it is shown the entire narrative preceding the question before trying to guess sentences in that narrative. The LLM's ability to predict sentences will be best when the preceding narrative is kept intact.

Of central interest is what might happen if the order of the sentences is shuffled. Shuffling sentences should disrupt the flow of ideas across the sentences, which may reduce the benefit of the narrative to help the LLM predict sentences in the original text. Shuffling the sentences should certainly reduce performance for individuals with narratives in which there is a coherent flow of ideas. Sentence shuffling might not reduce the benefits of the context when the sentences lack a coherent flow of ideas, as would be the case if each sentence was unrelated to the other sentences in a narrative. Thus, by investigating how an LLMs ability to generate sentences is affected by different kinds of context, we can determine the degree to which conceptual links between sentences are present. In disorganized narratives, the effect of sentence shuffling on prediction should be minimal.

Our work here seeks to discover any differences in which level of disruption (i.e., the conceptual/sentence level or the word level) affects CHR+ and CHR- more significantly. Given that LLMs generate predictions based on the content of the context they are given, we were able to take a feature engineering approach and manipulate this context to interrogate both idea-level and word-level disruptions in this CHR sample and these disruptions' effect on the model's ability to predict subjects' language. Broadly, because language may be better intact both at the word and idea levels in

those CHR-, we predict they would be more predictable by an LLM overall across these conditions. Further, we expected there to be differences in word-level findings, specifically that these will be less relevant to CHR- than CHR+, while idea-level disruptions may be more relevant to CHR-.

More specifically, we have six main predictions. 1) The LLM's ability to predict subjects' language will be best when the sentences of context provided to it are kept intact in contrast to when the context sentences are shuffled. This is expected because shuffling context sentences should result in a loss of information across sentences, reducing the LLM's predictive ability. 2) Critically, shuffling the sentences of context should have a greater impact on the LLM's ability to predict language from CHR- than CHR+. It is expected that the narratives of CHR- will have greater organization across sentences than those of CHR+, who exhibit greater disorganization in the thought processes. Thus, shuffling sentences should have a greater impact on the information contained across sentences in CHR+ than CHR-. 3) Shuffling the words within each sentence of context, but not the order of the sentences themselves, should reduce LLM's ability to predict language in both CHR+ and CHR-. Agrammaticisms are not a defining feature of psychosis, so disrupting connections between words should impact both groups (Radanovic, Sousa, Valiengo, Gattaz, & Forlenza, 2013). 4) While both groups should be affected by shuffling the words within sentences of context, this may have a greater impact on predicting sentences in CHR- than CHR+. This is because word shuffling might disrupt the flow of ideas within a sentence in addition to disrupting grammatical relations. Assuming that the flow of ideas within a sentence is more organized in CHR- than CHR+, then shuffling the words should have a greater impact on CHR- than CHR-+. 5) Assuming that loosening of associations at the word level leads to usual word choice, we propose that the LLM will have more difficulty predicting individual words in CHR+ language than CHR- language. 6) Based on prior literature by Shi and Wolff (2021), we predict that providing the LLM with some amount of context (rather than none at all) will help it to predict

subjects' language. We expect that this will be true more for CHR- than CHR+ because the context of CHR- contains more relevant information than the context of CHR+.

Methods

Participants

The dataset included 30 CHR individuals selected from a pool of participants in the second North American Prodromal Longitudinal Study (NAPLS-2) from Emory University. Seven of the participants developed psychosis (CHR+), and 23 demographically matched individuals did not (CHR-). Though NAPLS-2 was a consortium study with many participants, the current study could only include subjects for whom audio-video recordings of their interviews were available.

Measures and Protocol

Each participant received a diagnostic Structured Interview for Psychosis-Risk Syndromes (SIPS) to assess CHR status at their baseline visit. These were recorded and transcribed for analysis. Follow-up visits were conducted every six months for two years to assess diagnostic outcomes and the development of threshold psychotic illness.

Conversion or transition to psychosis was determined by the Presence of Psychotic Syndrome (POPS) criteria on the SIPS, which is a rating of '6' on any of the positive symptom items. An individual was classified as converting to psychosis if they reported clinical delusions, hallucinations, grandiosity, or thought/communication disorder for a minimum of one hour per day, four days per week, for at least one month. Symptoms with this level of frequency and severity meet DSM-IV criteria for a threshold psychotic disorder.

To assess presence of any other disorders or symptoms, subjects received the Structured Clinical Interview for DSM-IV (SCID-IV – Axis I Disorders) as well as several cognition tests, including the Wechsler Adult Intelligence Scale (WAIS) to assess intelligence quotient (IQ).

Demographic variables were also collected and not significantly different between CHR+ and CHR- at baseline (see **Table 1**).

All study participants were native English speakers and consented to have their interviews recorded. Exclusion criteria for all groups included the diagnosis of an existing threshold psychotic disorder, substance dependence, a neurological disorder, or an IQ <70 (for further details on exclusion criteria for NAPLS studies see Addington and colleagues (2015)).

	CHR-	**CHR+**	***p*-value**
Total N=30	**N = 23**	**N = 7**	
Age	21.496 ± 4.513	20.571 ± 5.623	0.294
Gender (% male)	47.82%	57.14%	0.666
Race			0.554
Caucasian	34.78%	14.29%	
Black	47.82%	57.14%	
Other	17.39%	28.57%	
Estimated IQ	103.363 ± 15.966	97.714 ± 13.124	0.520
Medication			0.109
Antipsychotics	0%	14.29%	
Antidepressants	14.35%	14.29%	

Table 1. Demographic data for CHR- and CHR+ in the training and test datasets.

Procedures

A unique feature of transformer models, including T0, is that they can generate sentences and paragraphs. We asked T0 to predict (or generate) participants' responses to the interviewer's questions, hereafter "target" sentences (see **Supplementary Materials** for a detailed description of LLM T0). First, we provided the model with varying amounts of context (ranging between 0-20 sentences directly preceding the target question) from the interview to ascertain the amount of context T0 required to make a prediction for the target sentence (see **Figure 1A**). The benefits of

context peaked when T0 was provided six sentences of context preceding the target sentence. Therefore, we adopted the use of six sentences of context for the conditions described below. Once the model generated target sentences, we compared them to what was produced by the participant.

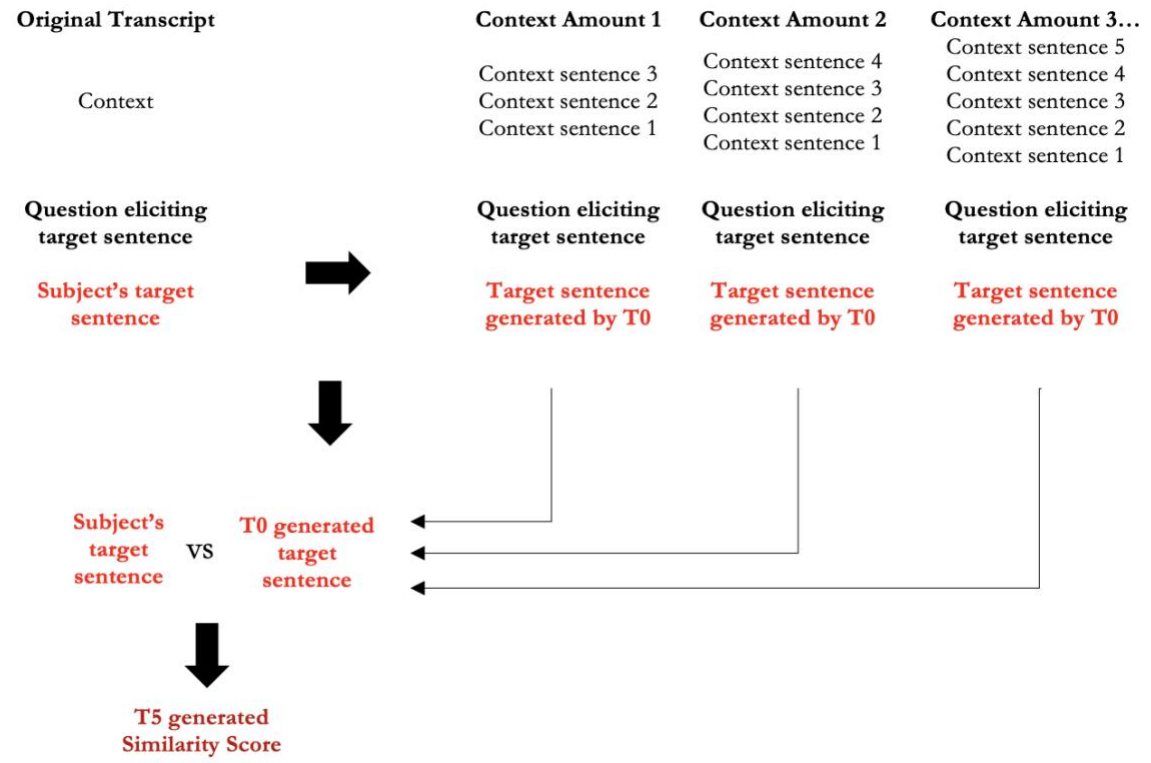
The degree of fit between the sentences produced by T0 and participants was determined by T5 similarity ratings (Raffel, et al. 2020). Similarity scores (0-to-5, where 0 = no similarity; 5 = identical) were produced by prompting T5 with a text string containing the keywords "stsb sentence1:" and "sentence2:", followed by sentences produced by the participant and T0, respectively. T0 was prompted to generate five guesses for each target sentence, with the highest similarity score retained for further analysis. T5's similarity ratings are based on a training set of 5749 human similarity ratings of sentences drawn from news headlines, question-answer forums, and image captures. T5's similarity ratings agree well with humans (r = 92.7), slightly exceeding the performance of individual humans on the STS-B GLUE Benchmark (Wang, et al., 2018).

Design

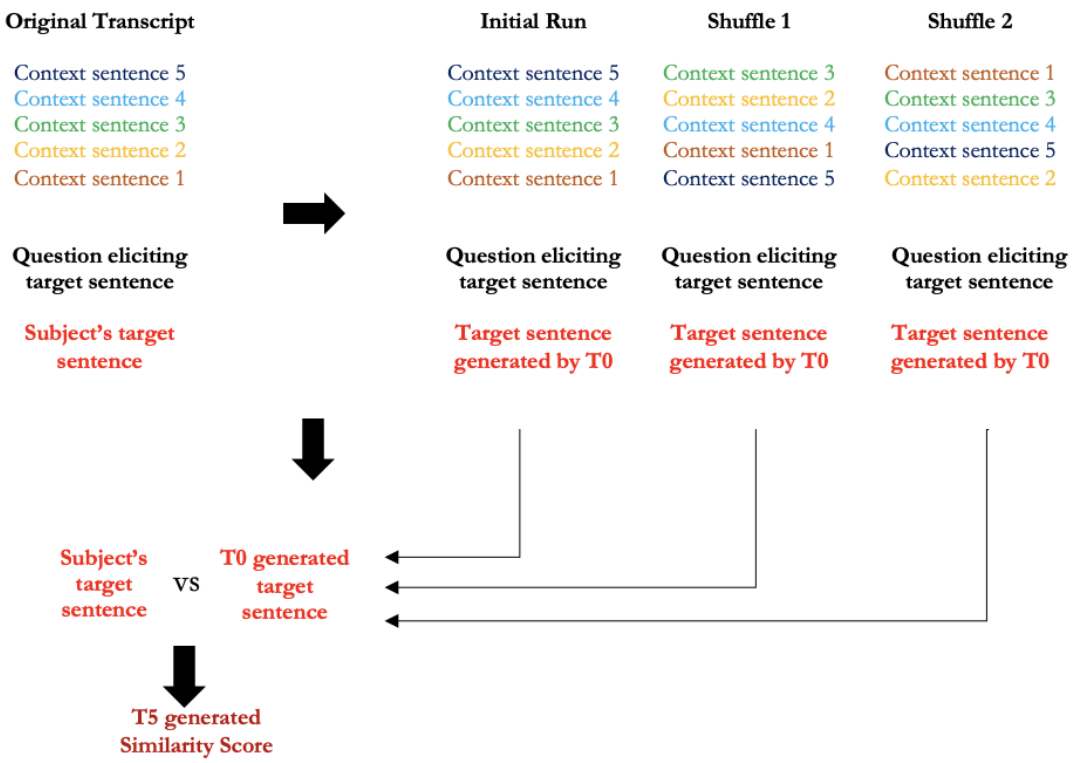
T0's prediction performance was evaluated across five conditions. 1) *Intact*. In the intact condition, the model attempted to predict the target sentence with the help of the six sentences preceding it, with the last sentence in the context being the question asked by the interviewer (see **Figure 1A)**. 2) *Context absent*. In the context absent condition, the model was prompted only with the question asked by the interviewer. 3) *Sentence shuffle*. In the sentence shuffle condition, the target sentence was preceded by the same six sentences as in the intact condition, except that the order of the sentence was shuffled, except for the question, which was kept as the last sentence in the context. The shuffle process was repeated ten times to establish reliable results (see **Figure 1B)**. 4) *Word shuffle*. In the word shuffle condition, sentence order was preserved, but the words in all of the sentences except the question sentence were shuffled. The shuffling process was repeated ten times to establish reliable results (see **Figure 1C)**. 5) *Word masking*. In the word masking condition, the

target sentence was preceded by three prior sentences, while each word in the target sentence was sequentially masked (see **Figure 1D)**. In this last condition, model performance was based on the percentage of time the model was able to guess the masked word.

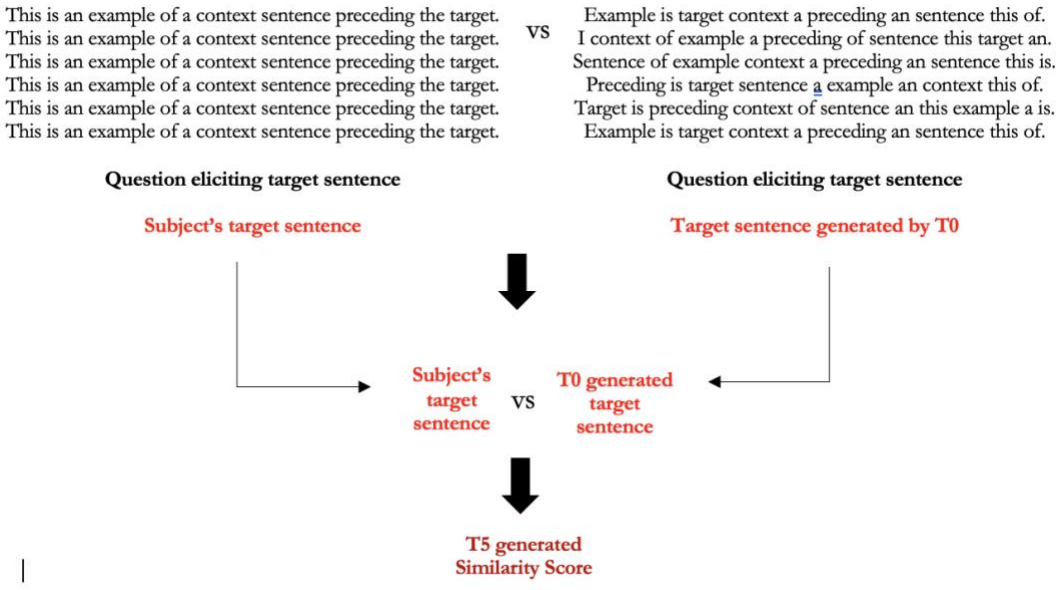
A.



B.


Original Transcript
Context sentence 5
Context sentence 4
Context sentence 3
Context sentence 2
Context sentence 1
Question eliciting target sentence
Subject's target sentence
Initial Run
Context sentence 5
Context sentence 4
Context sentence 3
Context sentence 2
Context sentence 1
Question eliciting target sentence
Target sentence generated by T0
Shuffle 1
Context sentence 3
Context sentence 2
Context sentence 4
Context sentence 1
Context sentence 5
Question eliciting target sentence
Target sentence generated by T0
Shuffle 2
Context sentence 1
Context sentence 3
Context sentence 4
Context sentence 5
Context sentence 2
Question eliciting target sentence
Target sentence generated by T0
Subject's target sentence
VS
T0 generated target sentence
T5 generated Similarity Score

C.


This is an example of a context sentence preceding the target.
This is an example of a context sentence preceding the target.
This is an example of a context sentence preceding the target.
This is an example of a context sentence preceding the target.
This is an example of a context sentence preceding the target.
This is an example of a context sentence preceding the target.
VS
Example is target context a preceding an sentence this of.
I context of example a preceding of sentence this target an.
Sentence of example context a preceding an sentence this is.
Preceding is target sentence a example an context this of.
Target is preceding context of sentence an this example a is.
Example is target context a preceding an sentence this of.
Question eliciting target sentence
Subject's target sentence
Question eliciting target sentence
Target sentence generated by T0
Subject's target sentence
VS
T0 generated target sentence
T5 generated Similarity Score

D.

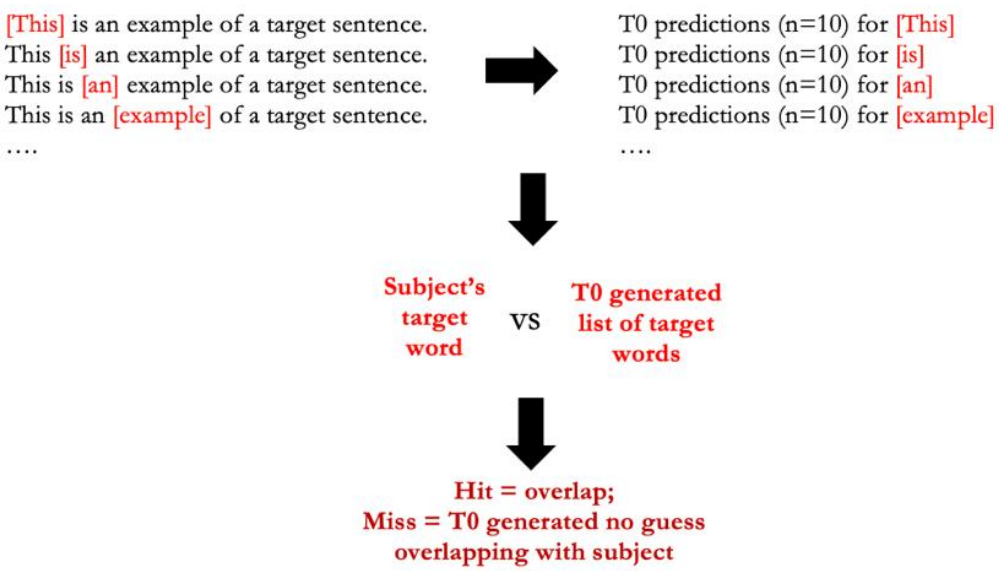


Figure 1. A. Each question posed to the question was answered by model T0 with varying amounts of context preceding the question. **B.** Six sentences of context preceding each question was then shuffled before T0 generated a response to the question. **C.** The words within each of 6 sentences of context directly preceding each question was shuffled before T0 generated a response. After each condition, T0's responses were then compared to the subject's utterance using a T5's similarity score. **D.** Each word in subjects' sentences from the transcripts was occluded iteratively and T0 generated 10 guesses for each one. Any match between the generated list and what was said by the subject was counted as a "hit."

Results

The results were largely as predicted. Most importantly, shuffling sentences had a bigger impact on T0's ability to predict target sentences for CHR- than CHR+. This result is consistent with theories suggesting that psychosis spectrum disorders involve non-local disturbances in the organization of ideas. In addition, T0's ability to predict target sentences from CHR+ were less affected by word-level shuffling than CHR-, indicating their intact language is affected at the word-level more so than CHR-.

Of central interest was whether sentence predictability would be reduced by sentence shuffling. If the model is picking up ideas across sentences and from how they build on one another, shuffling them should reduce the amount of information the model can use. We conducted a mixed

ANOVA on predictability between CHR+ and CHR- for the intact and sentence shuffling conditions using Greenhouse-Geisser correction for sphericity. The main effect of context did not reach significance ($F[1,28] = 0.16$, $p = .69$) but there was a significant interaction between conversion status and context type ($F[1,28] = 4.72$, $p = .038$; See **Figure 2A**).

We ran post-hoc pairwise t-tests to determine group-level difference between intact and sentence shuffle conditions. As expected, prediction was impacted by shuffling in CHR- as indicated by a trend-level significant difference between the intact and shuffling conditions ($t[22] = 2.6$, $p = .014$ (Bonferroni corrected α level to determine significance, $p < .05/4=.0125$)). Conversely, there was no significant difference in CHR+ uttered target sentences and the target sentences T0 generated between the intact and sentence shuffling conditions ($t[6] = -1.1$ $p = 0.83$). The fact that there was no change in predictability between intact and sentence shuffling conditions for CHR+ indicates context had no effect on predictability and that they may already exhibit "shuffling" in their intact speech.

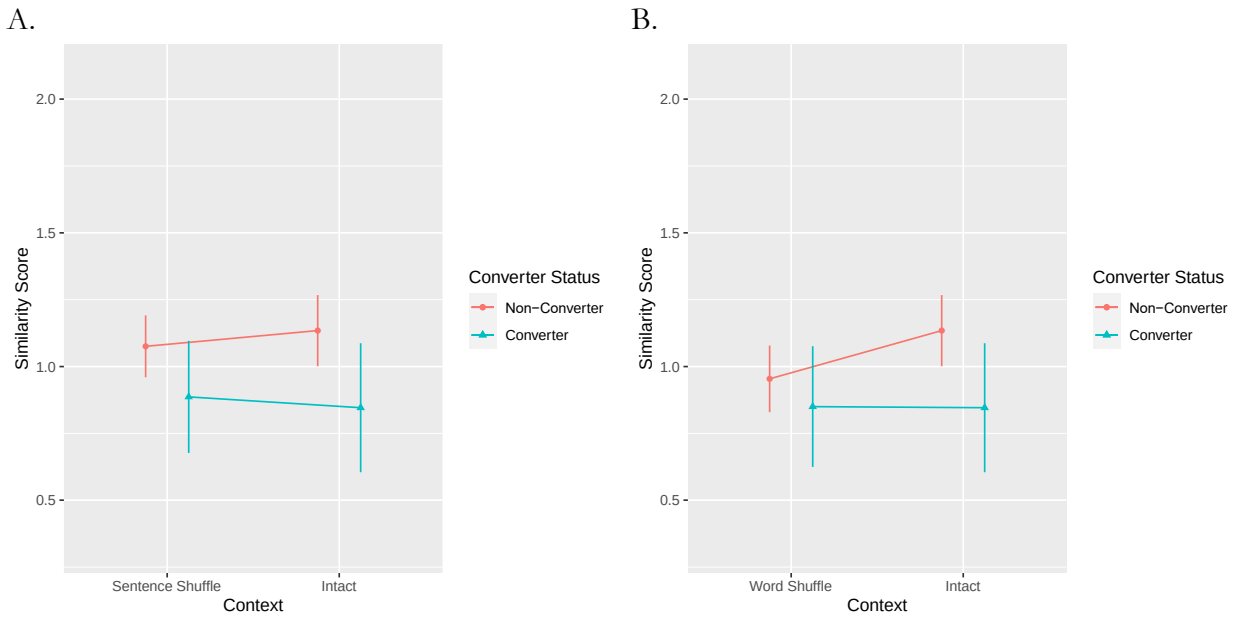


Figure 2. A. Interaction between sentence shuffling and intact context conditions by converter status **B.** Interaction between word shuffling and intact context conditions by converter status

Similarly, to sentence shuffling, we predicted that T0's performance would be impacted by shuffling the words in the context it was given. To test this, we conducted a mixed ANOVA on predictability between CHR+ and CHR- for the intact and word shuffling conditions using Greenhouse-Geisser correction for sphericity. There was a main effect of context ($F[1,28] = 7.77$, $p = .009$) and there was a significant interaction between conversion status and context condition ($F[1,28] = 8.53$, $p = .007$; See **Figure 2B**).

We ran post hoc pairwise t-tests to determine group-level differences between intact and sentence-level conditions. As expected, CHR- predictability differed significantly between the intact and the word shuffling conditions ($t[22] = 5.78$, $p = <.001$ (Bonferroni corrected α level to determine significance, $p < .05/4=.0125$)). However, not in line with our predictions, there was no significant difference between these conditions for CHR+ ($t[6] = -0.8$, $p = .53$). That is, surprisingly, shuffling the words within context sentences did not significantly reduce the model's ability to predict target sentences.

We also predicted that while both groups should be affected by shuffling the words within sentences of context, this may have a greater impact on predicting sentences for CHR- than CHR+. This is because word shuffling might disrupt the flow of ideas within a sentence in addition to disrupting grammatical relations. We calculated a difference score between predictability for intact and word-shuffled condition for each group. We then conducted an independent t-test of the difference scores, which indicated that predictability for CHR- was more impacted by word-shuffling than for CHR+ ($t[28] = 2.9$, $p = .007$).

Finally, as expected, T5 was less able to predict words in CHR+ sentences than CHR- sentences. This was true both when there was no context ($t[11] = 2.4$, $p = .036$) and when the model was given one sentence of directly preceding context ($t[20] = 2.5$, $p = .021$)

Other aspects of the results were also as predicted. First, a key assumption in these analyses is that T0's ability to predict sentences would benefit from context. As expected, when provided no context preceding each question eliciting the target sentence in the interview, there was no significant difference in how well T0 could predict responses to questions between CHR+ and CHR-. This was indicated by an insignificant t-test between CHR+ and CHR- based on their similarity scores between subject- and T0-generated responses ($t[17] = 1.36$, $p = 0.19$). When varying the amount of context that was provided to the model (see **Figure 1A**), we included conditions of 1-10 sentences as well as 20 sentences of context. Averaging across all these conditions, similarity scores were overall higher for CHR- than CHR+ ($t[10] = 10.45$, $p < .001$), supporting our prediction that the model performs better when it is given context on which to base a prediction. This finding also supported our prediction that CHR- would be overall more predictable than CHR+. This difference in predictability peaked when using six sentences of context ($t[12] = 2.37$, $p = 0.036$). However, in this condition of six intact sentences of context, the interaction between conversion status and context did not reach significance ($F [1, 28] = 2.31$, $p = 0.14$).

Discussion

Loosening of associations is a phenomenon associated with all stages of psychotic disorders (Kuperberg, 2010). It occurs at both the local level and non-local levels. To date, both of these levels have been examined largely by clinician observation and judgment. However, ratings rooted in observation of behavior are particularly vulnerable to subjectivity, and a scale as comprehensive as Andreasen's TLC is laborious to complete and requires training and clinical experience. Our work uses these core features of TD as a framework, and leverages LLMs to more objectively measure disorganization in CHR.

In this set of analyses, we took advantage of T0's generative nature and had it produce responses to questions that CHR subjects were posed during interviews. We manipulated several features of this process both at the level of sentences/ideas and words. By shuffling the sentences of context, we interfered with the flow of ideas from one sentence to another. Doing so kept the semantic and syntactic relationships between words intact while disrupting the conceptual organization of the subject's narrative. Notably, this diminished the predictability of CHR- but not CHR+, indicating that CHR+ may have a less linear or coherent flow of ideas to begin with. By then shuffling the words in sentences of intact context, we interfered with the relationships between words. As a byproduct of this drastic shuffling, we also interfered with the flow of ideas. This again reduced the predictability of CHR- but not CHR+. That is, when the words in the context were shuffled, predictability was negatively impacted compared to the sentence shuffling condition for CHR-, but not for CHR+ indicating that perhaps CHR- are less affected at the word level than CHR+. To further test word-level disruptions, we had T0 predict each word in each sentence the subject uttered during the interview for both groups. We found that CHR+ were significantly less predictable than CHR-, supporting that they are more disorganized at the word level than CHR-.

Perhaps the most striking finding is that shuffling context at both the local and non-local levels did not significantly impact an LLM's ability to predict language from CHR+. This aligns with our prediction that CHR+ are disorganized at the word level and significantly more so than CHR-. These findings may also explain some of the model's difficulty predicting the sentence shuffling condition for CHR+ because word-level shuffling also interferes with the conceptual organization of a narrative. If CHR+ speech is naturally disorganized or shuffled at the word-level, this would disrupt their flow of ideas across sentences.

Predictability for CHR-, however, was impacted by sentence shuffling and by word shuffling, which signifies their speech is more intact at the word and conceptual level. To our knowledge, this

is the first computational measurement of non-local disorganization akin to the derailments proposed by Kraepelin (1919) and tangentiality and derailment defined by Andreasen's (1986) TLC.

Using T0, we were able to conduct a more accurate and objective measurement of this central illness feature. This was made possible by a number of unique features of LLMs. Their generative nature allows us to go beyond local relationships in language that have been the basis of most computational work thus far, and their ability to produce whole sentences was a key component of these analyses.

Notably, LLMs are context-aware, meaning they can dynamically adjust the definition of words based on the context they appear in. This makes their predictions more accurate than context-free models that operate with a single definition for a word, regardless of context. Moreover, psychosis literature to date has focused on defining organization and coherence as high similarity between adjacent words and sentences, which is a highly restrictive definition of coherence. For example, when asked, *I'm going to get groceries – do you want to come to the store?* and someone responds, *No, I'm good,* they are producing a highly predictable response even though "good" may not be semantically similar to "groceries" or "store", the semantic content of the preceding question. Prediction uses similarity between sentences, but also allows for entailments and other features of language to contribute to performance. Therefore, using such a model allows us to adopt a less restrictive definition of coherence.

Research conducted using LLMs is also less vulnerable to overfitting than methods based in machine learning (ML) approaches. ML models are trained on one dataset and then tested on another sample. When using ML, there is a possibility that findings reflect more about the dataset they were trained on than the general phenomenon of interest. LLMs are trained on large amounts of language data so that they can understand and "use" language. Instead of fine-tuning a model to

our dataset, we were able to take a more experimental approach and manipulate aspects of our data to see their comparative effects on what the model generated.

This study provides evidence that it is possible to capture non-local disorganization across a narrative, but it was not imperative that we use T0 to do so. A number of LLMs can be used to do these calculations, making it a more versatile and robust approach than ML. Moreover, as LLMs are further trained and acquire an even deeper understanding of language, our ability to detect this non-local disorganization will likely improve.

While it is clear LLMs provide significant advantages over other computational approaches, there are also limitations to using them. For one, there is a baseline assumption that the language they generate that is very similar to "healthy" or normal speech. However, at this time there is no metric which we can use to definitively assert this claim other than evidence that they can write text that is perceived to be written by humans.

It is important to note that the shuffling method we employ here does not rely on T0 generating a sentence that would be highly predictable or analogous to "healthy" language, but instead on relative comparisons between intact and shuffled speech. That said, more work needs be done to understand if T0 uses context in the same way as humans. For example, LLMs and humans may differ in how much they weight contextual information closest to the target sentence for their predictions. In future studies, we can specifically manipulate the order in which sentences of context appear to the model to examine how this affects their ability to predict the target between these conditions.

Distinguishing local and non-local disorganization quickly and effectively could aid screening and treatment of TD in CHR. Local disorganization may indicate an elevated risk for conversion to psychosis, which is useful for early detection and improving prognostic outcome. It could also indicate disturbances associated with the production of language that requires intervention at a

young age. In contrast, people with non-local disturbances may benefit from social skills training that emphasizes effective communication and organization of ideas.

It will be important to confirm and expand these findings. A larger dataset, particularly with more CHR+, is needed to replicate these findings. Moreover, including healthy controls is an important next step in assessing the extent of disorganization in CHR+ and CHR-. It may be that CHR- are as predictable at the local level but display more non-local disorganization than healthy controls. Similarly, using these measures on longitudinal data will be important to examine if and how TD fluctuates with symptom changes. It is likely that local disorganization is a stable phenomenon throughout the developmental trajectory given that it can be used to predict conversion to psychosis before its onset. However, the non-local disturbance measured by T0 may vacillate in tandem with positive symptoms (i.e., loosening of ideas becomes more severe during mania).

Using different types of interviews is also crucial to supporting our findings. The interviews used in this study s were semi-structured interviews for which subjects were asked the same set of questions, in addition to differing follow-up questions to their responses. Semi-structured interviews provide continuity in the content subjects discuss, which can be beneficial for between subject and group comparisons. However, interviewers guide a subject's topic and may even interrupt subjects once critical information has been gathered, limiting the amount of non-local disturbance people exhibit. Using open-ended interviewing or description tasks allows subjects to model the full extent of their TD because of the limited structuring they are provided. We expect that our results will be robust to and even enhanced by these replication efforts.

References

Andreasen N. C. (1979). Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. *Archives of general psychiatry*, *36*(12), 1315–1321. https://doi.org/10.1001/archpsyc.1979.01780120045006

Andreasen N. C. (1979). Thought, language, and communication disorders. II. Diagnostic significance. *Archives of general psychiatry*, *36*(12), 1325–1330. https://doi.org/10.1001/archpsyc.1979.01780120055007

Andreasen, N. C. (1983). Scale for the assessment of negative symptoms (SANS The University of Iowa Press.

Andreasen, N. C. (1983). Scale for the assessment of positive symptoms (SAPS). The University of Iowa Press.

Andreasen N. C. (1986). Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia bulletin*, *12*(3), 473–482. https://doi.org/10.1093/schbul/12.3.473

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. npj Schizophrenia, 1(1), 1-7.

Bleuler, E. (1952). Dementia Praecox; or the Group of Schizophrenias (J. Zinkin, Trans.; D. C. Lewis, Foreword). International Universities Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., ... & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry, 17(1), 67-75.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophrenia research, 93(1-3), 304-316.

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of neurolinguistics*, *23*(3), 270-284.

Gordinier, S. W., & Docherty, N. M. (2001). Factor analysis of the Communication Disturbances Index. *Psychiatry research*, *101*(1), 55–62. https://doi.org/10.1016/s0165-1781(00)00239-0

Hart, M., & Lewine, R. R. (2017). Rethinking Thought Disorder. *Schizophrenia bulletin*, *43*(3), 514–522. https://doi.org/10.1093/schbul/sbx003

Johnston, M. H., & Holzman, P. S. (1979). Assessing schizophrenic thinking. Jossey-Bass Publishers.

Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. Schizophrenia Bulletin, 13(2), 261-276. https://doi.org/10.1093/schbul/13.2.261

Kraepelin, E. (1971). Dementia Praecox and Paraphrenia (R. Barclay Mary, Trans.; G. M. Robertson, Ed.). R. E. Krieger Pub. Co.

Kuperberg G. R. (2010). Language in schizophrenia Part 1: an Introduction. *Language and linguistics compass*, *4*(8), 576–589. https://doi.org/10.1111/j.1749-818X.2010.00216.x

Liddle, P., Ngan, E., Caissie, S., Anderson, C., Bates, A., Quested, D., . . . Weg, R. (2002). Thought and Language Index: An instrument for assessing thought and language in schizophrenia. *The British Journal of Psychiatry, 181*(4), 326-330. doi:10.1192/bjp.181.4.326

Minor, K. S., & Lysaker, P. H. (2014). Necessary, but not sufficient: links between neurocognition, social cognition, and metacognition in schizophrenia are moderated by disorganized symptoms. Schizophrenia research, 159(1), 198-204.

Minor, K. S., Marggraf, M. P., Davis, B. J., Luther, L., Vohs, J. L., Buck, K. D., & Lysaker, P. H. (2015). Conceptual disorganization weakens links in cognitive pathways: disentangling neurocognition, social cognition, and metacognition in schizophrenia. Schizophrenia Research, 169(1-3), 153-158.

Piantadosi, S. T. (2021). Modern language models refute Chomsky's approach to language. MIT Technology Review. https://www.technologyreview.com/2021/03/01/1020366/language-models-chomsky-artificial-intelligence-ai/

Radanovic, M., Sousa, R. T., Valiengo, L., Gattaz, W. F., & Forlenza, O. V. (2013). Formal Thought Disorder and language impairment in schizophrenia. Arquivos de neuro-psiquiatria, 71(1), 55–60. https://doi.org/10.1590/s0004-282x2012005000015

Rocca, P., Galderisi, S., Rossi, A., Bertolino, A., Rucci, P., Gibertoni, D., ... & Goracci, A. (2018). Disorganization and real-world functioning in schizophrenia: results from the multicenter study of the Italian Network for Research on Psychoses. Schizophrenia Research, 201, 105-112.

Roche, E., Creed, L., MacMahon, D., Brennan, D., & Clarke, M. (2015). The Epidemiology and Associated Phenomenology of Formal Thought Disorder: A Systematic Review. *Schizophrenia bulletin*, *41*(4), 951–962. https://doi.org/10.1093/schbul/sbu129

Saks, E. (2007). *The Center Cannot Hold: My journey through madness*. Hachette Books.

Shi, H., & Wolff, P. (2021). What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/0kr3t179

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for nature language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP, p. 353-355, Brussels, Belgium, Association for Computational Linguistics.

Woods, S. W., Walsh, B. C., Powers, A. R., & McGlashan, T. H. (2019). Reliability, validity, epidemiology, and cultural variation of the Structured Interview for Psychosis-risk Syndromes (SIPS) and the Scale Of Psychosis-risk Symptoms (SOPS). *Handbook of attenuated psychosis syndrome across cultures: International perspectives on early identification and intervention,* 85-113.

Supplementary materials:

Large Language Models (LLMs):

Language understanding was accomplished using the large language models (LLMs) T0 (T-zero) (Sahn et al., 2022) and T5 (Raffel, et al., 2020). T0 is based on T5 but with further training. While T0's inference capabilities are superior to T5's, we continued to use T5 for some aspects of this work because it is pre-trained to assess sentence similarities, which proved useful in choosing the T0 inferences that most closely matched those generated by the participants. Interestingly, T0's performance on several English natural language tasks is better than the well-known LLM GPT-3 (Brown, et al, 2020), while being 16 times smaller (Sanh et al., 2021).

Both T0 and T5 are composed of two main parts: a sequence of encoders and a sequence of decoders. In translation, the encoders "comprehend" sentences in one language (e.g., English), and the decoders generate text strings in another language (e.g., French). T0 and T5 extend the text-to-text paradigm to a wider range of natural language tasks, including next-sentence prediction, question answering, abstractive summarization, grammaticality assessment, and entailment/implication assessment (Raffel, et al., 2020; Sahn et al., 2022).

A novel feature of LLMs like T5 and T0 is that they have "self-attention." Self-attention allows these networks to dynamically adjust the meaning of words to fit a context. For example, LLMs can adjust the meaning of the word *run* to capture its different meanings in different contexts, as in "to run a marathon," "to run a campaign," or "to run a wire between the poles." LLMs acute sensitivity to context enables them to predict missing words and sentences with uncanny accuracy, even when the local transition probabilities between the words are low. The ability to predict non-typical words and sentences is evidenced by their ability to generate sentences similar to Chomsky's "Colorless green ideas sleep furiously", as in "Purple fluffy clouds drame wildly" (Piantadosi, 2023). While LLMs can seemingly comprehend and produce language like humans, their impressive

performance does not entail that they process language in the same way as humans. *How* LLMs process information may be quite different from humans. Here we merely use LLMs to measure information content and investigate how information content might differ across individuals. We do not use them as models of human processing.

Subsetting CHR- :

To ensure that effects presented in the results were not due to a small CHR+ sample size, we subset a group of seven CHR- and repeated the analyses described above. Like in the larger set, there is a significant difference between CHR+ and CHR- for the no context condition, but there was only a trend difference between CHR+ and CHR- in the condition containing 6 sentences of intact context ($t[12] = 1.8$ $p = 0.09$), and no group differences in the sentence shuffle condition ($t[12] = .98$ $p = 0.35$) or the word shuffle condition ($t[12] = 1.49$ $p = 0.16$). These results match the ones above for the entire group of CHR-.

Within the subset of CHR-, there was a significant difference between the no context and intact conditions ($t[6] = 3.22$ $p = 0.009$). There was no significant difference between intact and sentence shuffle conditions ($t[6] = 0.6$ $p = 0.29$), however, there was still a significant difference between the intact and the word level shuffle conditions ($t[6] = 3.02$ $p = 0.01$). This indicates that prediction was impacted most when the words were shuffled and that their speech is relatively intact at the word level remains even in this underpowered subsample.